# 一种基于词义和词频的向量空间模型改进方法 *

邓晓衡†，杨子荣，关培源

(中南大学 软件学院，长沙 410075)

**摘　要：**向量空间模型(VSM)是一种使用特征向量对文本进行建模的方法，广泛应用于文本分类、模式识别等领域。但文本内容较多时，传统的 VSM 建模可能产生维数爆炸现象，效率低下且难以保证分类效果。针对 VSM 高维现象，提出一种利用词义和词频降低文本建模维度的方法，以提高效率和准确度。提出一种多义词判别优化的同义词聚类方法，结合上下文判别多义词的词义后，根据特征项词义相似度进行加权，合并词义相近的特征项。新方法使特征向量维度大大降低，多义词判别提高了文章特征提取的准确性。与其他文本特征提取和文本分类方法进行比较，结果表明，该算法在效率和准确度上有明显提高。

**关键词：**文本分类；特征选择；卡方分布；向量空间模型

**中图分类号：**TP391.1　　doi: 10.3969/j.issn.1001-3695.2017.12.0752

## Method based on word meaning and word frequency to improve vector space model

Deng Xiaoheng†, Yang Zirong, Guan Peiyuan

(*School of Software*, *Central South University*, *Changsha* 410075, *China*)

**Abstract:** Vector space Model (VSM) is a method of modeling text using Eigenvector, which is widely used in the fields of text categorization and pattern recognition. But when the text content is more, the traditional VSM model may produce the dimension explosion phenomenon, the efficiency is low and the classification effect is difficult to guarantee. Aiming at the phenomenon of VSM, this paper proposes a method to reduce the dimension of text modeling by means of word meaning and frequency, in order to improve efficiency and accuracy. In this paper, we propose a synonym clustering method for polysemy discriminant optimization, combining with the context distinguishing word meaning, weighted by the similarity of the word meaning, and merging the feature items with similar meanings. The new method has greatly reduced the dimension of eigenvector, and polysemy has improved the accuracy of feature extraction. Compared with other text feature extraction and text categorization methods, the results show that the algorithm has a significant improvement in efficiency and accuracy.

**Key words:** text categorization; feature selection; chi-square; vector space model



## 0　引言

文本分类是考虑文本的属性与各个类别之间的匹配度来进行划分的过程 [1]。文本是自然语言处理的一个重要研究方向，在信息精准推送、信息过滤、网络传输优化等方向有极高的应用价值，被应用在广告精准推送、门户网站新闻筛选、购物平台精准推荐、社会话题挖掘、舆情分析、流感疫情监控等方面，具有非常重要的现实意义。

文本的 VSM 建模是将文本中词视为文本的特征项，根据一定规则将文本建成一个特征项的向量，但该方法会带来维数过高和数据稀疏性问题。过高的维数需要巨大的计算量，数据稀疏性问题意味着大量的无用计算，因此，文本降维是文本分类性能的瓶颈问题。通常使用特征选择方法来对文本进行降维。降低维数的一个主要方法就是特征选择，即根据词频或类别匹配度等信息评估最能表征文本的 p 个特征项，以此代表文本关键特征。

文本分类最重要的过程是特征选择，目前主要的文本特征选择方法有卡方检验[2]（CHI）、信息增益[3]（IG）、文档频次（DF）等。其中，卡方检验和互信息都表示文档主题与特定类别之间的相关性，CHI 值或 MI 与文档特征与特定类别呈正相关关系。以上几种文本特征提取方法没有绝对最优，在不同场合下有不同的表现效果。CHI 特征提取效果较好，相比其他方法计算代价更高。对于英文文本的特征提取，CHI 和 IG 的效果其他方法更好。在中文文本特征提取中，CHI 的效果最好，其次是 IG。

收稿日期：2017-12-01；修回日期：2018-01-24　　基金项目：中南大学研究生创新基金资助项目（2017zzts732）

作者简介：邓晓衡（1974-），男（通信作者），湖南省衡阳人，教授，主要研究方向为边缘计算、无线网络、大数据、分布式系统（dxh@csu.edu.cn）；杨子荣（1992-），男，安徽安庆人，硕士，主要研究方向为无线网络、模式识别、计算机视觉；关培源（1987-），男，博士，主要研究方向为边缘计算、群体智能感知、机会网络.

针对 CHI 模型的改进研究得到许多学者的关注。

现有文本分类模型的一个难点在于，如果要提高分类准确度，则维度不能太低；如果维度太高，又会大大降低分类效率；特征项的权值是基于频率来计算的，没有考虑语义和特征项之间的相关度。因此，为降低模型维度和权证项权值准确性，文本通过同义词词表对特征项进行同义词合并，在降低维度的同时提高关键特征项的权值，增强了特征项选择的准确性。此外，针对 CHI 方法对于低频次过于敏感的问题，使用词频对特征项权值进行优化，使模型达到更好的分类效果。针对同义词聚类时一词多义问题，提出基于多义词义项判别的同义词聚类优化方法。

## 1 相关工作

### 1.1 文本分类

20 世纪 60 年代，Salton 等人首次提出用向量空间的思想对文本建模，引入索引向量表征文本特征和属性，其良好的数学性质大大提高计算效率和准确性，广泛应用于文本分类、信息索引等领域。1997 年，Joachims 等人将支持向量机[4]引入 VSM 模型，提高了分类准确性。 2002 年， Chieu 等人将最大熵[5]引入文本分类，取得良好的效果。2005 年，Gutptal 等人将粗糙集[6]方法引入 N 元分类，大大减少分类模型训练时间。2005 年，Hirsch 等人将遗传算法引入文本分类，使用 TD-IDF 进行特征选择，取得良好分类效果。2006 年，Arunasalam 等人将关联规则方法引入文本分类，解决了文本分类时存在的类别不平衡的问题。2009 年，Yi 等人将因马尔可夫模型[7]引入医学文本分类，实现了医学领域不同分支的文本分类。

在文本特征项词性和词义方面，1997 年 Belhumeur 等人提出将短文本语义相似度[8,19]引入特征项过滤，将含义相似或相近的短语加权以提高文本分类准确度。2002 年 Yang 等人利用普林斯顿大学开发的英文语料库 HowNet 和 WordNet，进行词义联系在文本特征提取方面的研究，但是时间和空间复杂度都较高。2009 年，Gad 等人使用词义关系[10-12]优化了文本内特征项 TF 值，取得良好效果。

文本建模后，需要对其对其进行特征选择[13]，一种常用的方法是 CHI，CHI 基于概率统计模型，有良好的数学性，便于计算和分析。但是 CHI 也有一些缺点，如没有考虑文本的差异性，没有考虑特征项词义词性，对低频次过于敏感等问题。针对 CHI 的缺点，学术界提取许多改进方法。Li 等人考虑了不同分类中的文本差异性问题，提出类别权重因子，针对特定的类对模型进行优化。裴英博[14]等通过考察类别文本数，分析分散度和集中度对建模的加权影响，提高在语料库各类别文本数不均时的建模准确度。熊忠阳等人[15]针对文本类别库中特征分配不均匀等问题，将文本频数、特征分散度、集中度等参数引入 CHI 模型，提高了模型的分类准确性。王光等人[16]结合 CHI 和 IG 两种方法的特点，提出 CHI-IG 的特征选择方法，利用两种方法互补，提高模型稳定性和性能。以上方法考虑了建模因子和特征值权重对分类的影响，但是没有考虑特征分布差异性。邱云飞等人[17]提出一种改进的卡方函数，新增了三个参数用以更新特征值的权重，使得待选特征项更多地分布在某一类中。肖婷等将文本内的特征项频次引入模型加权，并将类内的正确度作用模型一个重要指标，优化 CHI 建模时低频次权重过高的问题。以上方法考虑的特征项的频次问题和分布差异问题，但是没有考虑正负相关性问题。Messad 等人将信息增益和文本频率结合到 CHI 中，提出三种方法组合的特征选择方案，弥补 CHI 的不足。Galavoti 等人提出一种特征项和类别正负相关性的方法，强调文本特征项对于分类起的作用。以上方法在特征项的词性上和传统 CHI 一样，都没有考虑特征项和特征项之间的相关性，而是把每个特征项都看做彼此独立的的单位，进从数学概率和建模方法上提出改进，对特征项词义方面并没有关注。

### 1.2 基于 VSM 文本分类模型原理及不足

人类能阅读抽象的文本信息，综合文本前后文的关系和语义逻辑，基于理解的方法找到文本的关键特征，从而找到其所属类别。但是计算机无法想人类一样理解文本，为使计算机也能对文本进行特征选择和分类，将文本进行分词后，统计其词频信息，根据文本词频和语料库词频的相关度找出一组特征词向量，再以此进行分类等操作。将抽象的文本拆分为词频的统计数据的过程即为 VSM 的核心思想。在 VSM 模型中，文本表征为一组特征项的集合，每一个特征项都有权值信息，表征该特征项的重要程度。TF-IDF 是用来估计一个词对某个文档集中的某分文档或整个语料库的重要程度，用来表示词的重要性，与该词在整个语料库中出现频率成反比，与该词在指定文档中出现次数成正比。

将文本分词建模后，其维数往往很大，需要对文本进行降维处理。使用 VSM 对文本进行分类前，需要对文本进行预处理，滤除无用信息，包括对文本进行格式化、分词、去停用词等。特征向量是根据一定规则提取文本特征集中的一部分特征项来表征文本内容，文本特征选择方法中最重要的是评估函数，对特征项的重要性进行评估，然后根据重要性从大到小排序，选择前 p 个作文文本的特征子集，达到降维的目的。

CHI 是一种经典特征评估函数。CHI 模型表述为式(1)：对于特征项 $t_k$，如果某一类的文本集 $c_j$ 中含有该项的文本数比例很大，其他类文本集中含有含项的文本数比例很小，则特征项 $t_k$ 对类别 $c_j$ 有越强表征能力。假设有特征集 $T=\{t_k \mid k=1,2,3,\ldots,m\}$，$c_j$ 为第 j 类文本集，该类别中的文本总数为 $n_{c_j}$，使用 CHI 来度量特征项 $t_k$ 和类别 $c_j$ 之间的相关性。

$$\chi^2(t_k,c_j)=\frac{n(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

其中：$t_k$ 表示特征项，$c_j$ 为第 j 类文本集，A 是 $c_j$ 类中含有 $t_k$ 的文本数，B 是语料库中 $c_j$ 外含有 $t_k$ 的文本数，C 是 $c_j$ 类中不含 $t_k$ 的文本数，D 是 $c_j$ 外不含 $t_k$ 的文本数，n 是语料库中文本数的总和。可知，(A+C)为类 c 的文本总数 $n_{c_j}$，(B+D)为语料库中类 $c_j$ 外的文本总数，这两个值都为常数，因此式（1）可简化

为

$$\chi^2\left(t_k,c_j\right)=\frac{\left(AD-CB\right)^2}{\left(A+B\right)\left(C+D\right)} \tag{2}$$

若 $t_k$ 和 $c_j$ 相互独立，则 $\chi^2(t_k,c_j)=0$，该值越大，则说明 $t_k$ 和 $c_j$ 相关性越强。

特征项的权值定为

$$\chi^2\left(t_k\right)=\max_{1\le j\le r}\left\{\chi^2\left(t_k,c_j\right)\right\} \tag{3}$$

进行特征项选择时，选择权值最大的前 p 个特征项来表征文本。

使用以上方法进行分本分类时可能出现的问题有：没有考察特征项的词性，认为特征项之间彼此独立，但实际情况中往往出现多个词代表一个含义的情况，针对这种情况，本文提出一种使用近义词来对特征项筛选和合并的方法。此外，传统 CHI 方法对低频次非常敏感，难以对低频次赋予正确的权值，当一个语料中的低频次在某文本中出现较多时，该低频次的权值就很大，而往往文本并非要表达该低频词，阵地该问题，本文提出一种根据词频来对特征项权值进行优化的方法，以避免低频次敏感的问题。

## 2　基于词义和词频的 VSM 模型改进方法

### 2.1　基于特征词词义的特征向量优化

汉语语言体系博大精深，同一种意思往往有多种表达方法，如“土豆”和“马铃薯”代表同一个意思，两者可以互换；“开心”和“愉快”代表同一类心情，但两者表示高兴的程度不一样，在某种程度下可以互换；“火车”和“动车”属于同一类交通工具，往往不能互换，但是有很大的相关性。而同一个意思往往也有很多种表达方式，如表达一个人很悲伤，可以用“悲痛欲绝”，也可以用“愁容满面”，只是两者表示的程度不一样。词语和词类是多对多的关系，如图 1 所示。

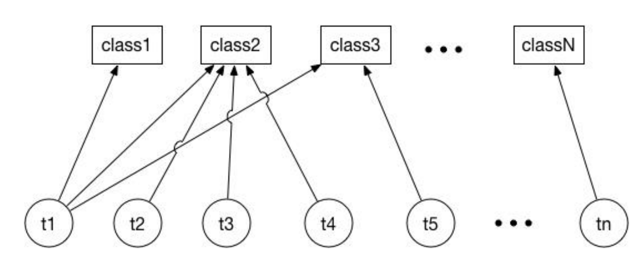



图 1　词与词类的多对多关系

中文的字词通常以读音为顺序进行编排，如《新华字典》《辞源》《辞海》等，近代将中文字词以词义进行统一编排的有 1983 年梅家驹版《同义词词林》。进入现代后，中文语言体系发生了很大变化，新词不断增加，旧词不断淘汰。哈工大信息检索实验室对中文词汇重新进行了一次系统编排，收录近 7 万条词汇，参照生物分类学方式将词汇按 5 个等级进行分类，如表 1 所示。文本对第 5 级进行合并和过滤。

表 1　词语汇编表

| 编码 | 符号举例 | 符号性质 | 级别 |
|---|---|---|---|
| 1 | A | 大类 | 第 1 级 |
| 2 | c | 中类 | 第 2 级 |
| 3 | 3 | 小类 | 第 3 级 |
| 4 | 6 | 小类 | 第 3 级 |
| 5 | B | 词群 | 第 4 级 |
| 6 | 1 | 原子词群 | 第 5 级 |
| 7 | 3 | 原子此群 | 第 5 级 |
| 8 | =%&@ | | |

在文本 Da 中，假如特征项 $t_i$ 和 $t_j$ 的含义完全相同，Da 中特征项 $t_i$ 的个数为 $n_i$，$t_j$ 的特殊为 $n_j$，则可将文本中全部的 $t_j$ 替换为 $t_i$，此时 $t_i$ 的个数为 $n_i+n_j$，重新计算 $t_i$ 的 TF-IDF 值，去除特征向量中的特征项 $t_j$，新的特征向量维数减 1。

对于含义相近但不能完全互换的词，可以根据其相似度进行加权。如合并第 5 级，则对第 4 级的词类统一编号，设词类总数为 N，设词类特征向量 $S(t_k)=\{class1:w_{k_1},class2:w_{k_2},\dots,classn:w_{k_n},\}$ 的维度为 N。对于文本 Da 的特征向量 $va=\{t1:w_{a_1},t2:w_{a_2},\dots,tn:w_{a_n},\}$ 中的特征项 $t_k$，如果它同时属于多个词类，则将其对应词类置为 1，其他类置为 0，如此将 $t_k$ 映射为词类向量 $t_k$。

则两个特征项 $t_i$ 和 $t_j$ 的相似度可用余弦相似度计算，如式（4）所示。

$$sim\left(S\left(t_i\right),S\left(t_j\right)\right)=\frac{\sum_{p=1}^{N}w_{i_p}\times w_{j_p}}{\sqrt{\sum_{p=1}^{N}{w_{i_p}}^2}\times\sqrt{\sum_{p=1}^{N}{w_{j_p}}^2}} \tag{4}$$

由于一篇文本的所属词类往往很多，用以上方式表述需要大量的计算。根据词和词类的对应特性，将词表述为词类的集合。如词 $t_i$ 同时属于 class1、class7、class15 这三个词类，则可将 $t_i$ 表述为 $C(t_i)$，$C(t_i)=\{class1,class7,class15\}$，则两个词的相似度可简化为

$$sim\left(S\left(t_i\right),S\left(t_j\right)\right)=\frac{\left|C\left(t_i\right)\cap C\left(t_j\right)\right|}{\sqrt{\left|C\left(t_i\right)\right|}\times\sqrt{\left|C\left(t_j\right)\right|}} \tag{5}$$

确定了两个特征项的相似度之后，则和对相似度高的特征项进行加权合并。假设特征项 $t_i$ 和 $t_k$ 的相似度为 $sim\left(S(t_i),S(t_j)\right)$，$t_i$ 的 TF-IDF 值较大，$t_j$ 的 TF-IDF 值较小，说明文本中该词义主要表述为 $t_i$。设文本中 $t_i$ 的个数为 $n_i$，$t_j$ 的文本个数为 $n_j$，则可将 $t_i$ 和 $t_j$ 合并为 $t_i$，新的 $t_i$ 值为 $n_i+sim\left(S(t_i),S(t_j)\right)\times n_j$。

根据《同义词词林扩展版》合并掉所有第 5 级的所有同义

词近义词，得到新的向量，为优化后的特征向量。

### 2.2 基于特征词词频率的低频词敏感优化方法

特征对类别的表征能力体现在两个方面，理想情况下，最能表征一个类别的特征项应该在该类语料库中大量出现，在其他类的则很少出现，反映到 CHI 的模型中，即类 $c_j$ 中含有特征项 $t_k$ 的文本数越多（A 值越大），在其他类别中含有特征项 $t_k$ 的文本书越少（B 值越小），则 $t_k$ 对类 $c_j$ 的表征能力越强。

分析 CHI 公式可知，$n_{c_j}$ 是类 $c_j$ 中文本总数，是常数，n 是语料库中的文本总数，也是常数，因此 A 越大则 C 越小，A/C 就越大，B 越小则 D 越大，B/D 就越小。根据 CHI 的思想，应寻找到 A/C＞B/D 特征项，即 AD-BC>0，(AD-BC)的值越大说明其表征类别的能力就越强。

但是，从 CHI 的公式中发现，如果某特征项的(BC-AD)值较大，该特征项也会被选中，对应情况是在其他类中出现较多，而在 $c_j$ 中出现概率较小的特征项，即 $c_j$ 中的低频次。CHI 的模型使得其对低频次敏感，则文本中的低频次往往不应成为文本的特征，甚至应当作为噪声去除。

以上两种情况分别为特征项和词类呈现的正相关性和负相关性。若文本类别 $c_j$ 中的特征项 $t_k$ 在该类中普遍出现，在其他词类中却很少见，说明 $t_k$ 和 $c_j$ 有强烈的正相关关系，待分类文本中如果该特征项大量出现，则可认为该文本与类别 $c_j$ 有很大的相关性，称为正相关。反之，如果 $t_k$ 在 $c_j$ 中几乎不出现，在其他类别出现次数较多，当文本中大量出现 $t_k$，则说明待分类文本有很大可能不属于类 $c_j$，称为负相关。

对于频率特别低的特征项，往往在文本处理阶段就会将其作为噪声去除，然而有一些中频词，往往具有一定低频词的属性，对类别 $c_j$ 可能呈现负相关性，但是 CHI 值却比较大。为了提高文本分类模型的准确度，将特征项对类别 $c_j$ 的影响能力分为正相关性和负相关性两个类别分别考虑。

$$\chi^2(t_k,c_j)^+ = \frac{(AD-CB)^2}{(A+B)(C+D)}, AD-CB>0 \tag{6}$$

$$\chi^2(t_k,c_j)^- = \frac{(AD-CB)^2}{(A+B)(C+D)}, AD-CB<0 \tag{7}$$

优化后的 CHI 公式为

$$\chi^2(t_k) = \alpha \times \sum_{1\le j\le r} \chi^2(t_k,c_j)^+ + (1-\alpha)\times \sum_{1\le j\le r}\chi^2(t_k,c_j)^- \tag{8}$$

$\alpha \in (0.5,1)$,是调节正负相关度比重的参数,实验部分将其取值为 0.8。

CHI 公式中，A,B,C,D 都是以文本数为单位进行计算的，没有考虑文档内频次。假定类别 c 中含有特征项 $t_i$ 的文本数为 $A_i$，含有特征项 $t_i$ 的文本数位 $A_j$，若 $A_i$ 与 $A_j$ 相等，则 CHI 认为特征项 $t_i$ 和 $t_j$ 对类别 c 的表征能力相同。但是实际情况中，如果文本内 $t_i$ 的频次高于 $t_j$，应当认为 $t_i$ 比 $t_j$ 对类 c 有更强的表征能力，但是 CHI 公式无法表征出这种差异性。

反之，一些文档频次不是很高，但是文本内频次很高的特征项却可能会被滤除，或者低估其权值。为此，应该考虑特征项 $t_k$ 在类 $c_j$ 中的所有文本内的频次。记 $tf_{ik}(t_k)$ 为特征项 $t_k$ 在类别文本 di 中出现的频次，则特征项 $t_k$ 在在类别 $c_j$ 中的出现频次为

$$tf_{jik} = \sum_{k=1}^{j} tf_{ik} \tag{9}$$

为使得作为优化参数代入公式，对其进行归一化，记优化权重因子 $\lambda_k$ 如式（10）所示。

$$\lambda_k = \sum_{j=1}^{r} \frac{tf_{jik}}{\sqrt{\sum_{k=1}^{m} tf_{jik}^2}} \tag{10}$$

将 CHI 公式更新为

$$\chi^2(t_k) = \lambda_k \times [\alpha \times \sum_{1\le j\le r}\chi^2(t_k,c_j)^+ + (1-\alpha)\times \sum_{1\le j\le r}\chi^2(t_k,c_j)^-] \tag{11}$$

此时，改进后的 CHI 模型降低了低频次的影响，增强了文档内频次高的特征项的权重。

### 2.3 基于多义词义项判别的同义词聚类优化方法

自然语言处理中，往往要面对各种消歧问题，如注音歧义、分词歧义、词义歧义、语用歧义等。其中，词义消歧往往需要针对特定的上下文来选择合适的含义，人类在理解语言时也时常会面临此类问题，如“德州扑克”和“德州扒鸡”中的“德州”指的是两个不同的地方，“一袋苹果”和“苹果手机”中的“苹果”指的是两种不同的东西。在自然语言处理中，同义词和多义词是一个普遍现象，一个词语可能同时属于多个词群，一个词群也包含多个词语，如图 2 所示。针对一词多义情况，如果不能正确的找出其含义，往往会对分类结果造成很大影响，因此，在使用同义词降维前，找到多义词和正确义项非常重要。

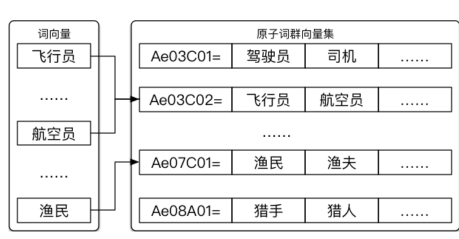



图 2　词语与词群的关系

1954 年 Harris 年提出分布式假说，认为“上下文相似的单词，语义也相似”，Firth 与 1957 年对该假说进行进一步阐述，认为“词的语义由其上下文决定”。一般认为上下文指定一个特定的语境，词语在相似的语境下的语义一般相似，通过计算上下文语境的相似度来对多义词词义进行标注，可减少同义词聚类时一词多义带来的影响。使用上下文对词语含义进行预测的方法称为 single sense word vector，一个单词对应一个词向量，没有考虑一词多义，结果即平均化的结果。该方法虽然不符合直观感受，在实际应用中却有较高的准确性和可靠性。针对相

似的上下文中，多义词不同含义的情况，上述方法并不全面，如“小明买了一袋苹果”和“小明买了一部苹果”，这两个句子上下文相似度非常高，但是“苹果”在两个句子中表示的含义完全不同。针对这种情况，有两种方法，第一，增加二元组中上下文词语集合的维度，使上下文包含的内容更多，含义更明确。第二，使用 multiple sense word vector，用来解决一词多义问题，计算在特定的上下文中，多义词的每个含义出现的概率，取最大条件概率为判别结果，即“一袋苹果”中，“苹果”是水果的概率大于“苹果”是手机的概率。具体过程为，建立特征项和上下文集合的二元组 $(s_i,V_{context})$，其中 $V_{context}=(v_1,v_2,\ldots,v_n)$ 表示 $s_i$ 的上下文词语集合，计算在特定上下文下多义词每个义项出现的概率，取最大条件概率的词义为判别结果。公式表示为

$$\max_{s_i\in S} p(s_i)\prod_{v_k\in V}^{n} p(v_k\mid s_i) \tag{12}$$

S 表示义项集合，$v_k$ 为上下文中的某一个词，该公式中的先验概率和条件概率无法通过语料库直接计算，而是使用同义词典中不同义项所属的原子词群在语料库中的分布估算。当语料库足够大时，通过统计的方法，可得到各个词义和上下文的组合关系，从而获得在特定上下文中多义词的词义。

本文在同义词聚类时，新增一个词义判别过程，为多义词找到正确含义。根据以上假说，使用上下文语境的方法来判别多义词的义项，具体实现过程为，建立一个 M*N 的矩阵，M 表示词义数，N 表示语料库中词语个数。为方便代码实现，使用一个 M 维向量存储同义词词典原子词群的编码，使用一个 N 维向量存储语料库中的词语。

算法过程如下：

a)根据《同义词词典》得到《多义词词典》

b)根据《多义词词典》得到编码向量

c)根据词语搭配库产生得到配对组合

d)得到多义词义项判别矩阵

整体算法框图如图 3 所示。

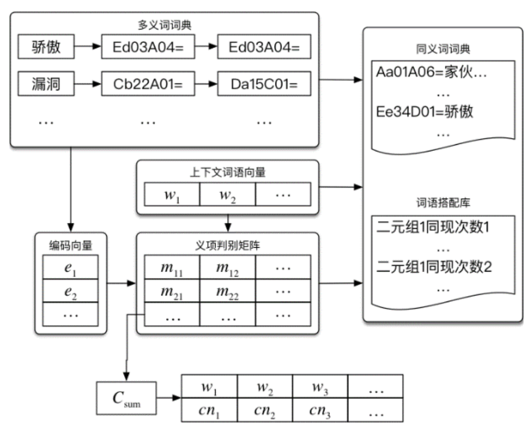



图 3 多义词义项判别算法框图

算法实现关键代码如下所示：

算法：多义词义项判别方法

输入：同义词词典（synonymicon），多义词词典（polysemantic dictionary），词语组合库（wordSetR），编码向量（Code Vector），词语组合向量（Word Vector）

输出：多义词义项判别矩阵（Matrix）

```
for(int i=0;i!=R.size();i++){
  //A 为搭配左词，B 为搭配右词
  if(PolyDic.find(A)){//A 为多义词
    if(WordVec.find(B)){//B 在上下文中
      while(P++){
        //返回编码向量中的索引
        i=GetIndex(CodeVec,*P++);
        //返回 B 在上下文中的索引
        j=GetIndex(WordVec,B);
        if(PolyDic.find(B)){//B 为多义词
        //H 为 B 的义项数
          Matrix[i][j]+=count/H;
        else
          Matrix[i][j]+=count;
        }
      }
      if(PolyDic.find(B)){//B 为多义词
        P=GetHead(B);
        if(WordVec.find(A)){//A 在上下文中
          while(P++){
          //返回编码索引
          i=GetIndex(CodeVec,*P);
          //返回 A 在上下文中的索引
          j=GetIndex(WordVec,A);
          if(PolyDic.find(A))
            //L 为 A 的义项数
            Matrix[i][j]+=count/L;
          else
            Matrix[i][j]+=count;
          }
        }
      }
    }
  }
}
```

在得到多义词的义项判别矩阵之后，利用以下公式估算先验概率 $p(s_i)$ 和条件概率 $p(v_k\mid s_i)$，公式如下：

$$p(s_i)=\frac{c(s_i)}{\sum_{i=1}^{m}c(s_i)} \tag{13}$$

$$p(v_k\mid s_i)=\frac{c(s_i,v_k)}{c(s_i)} \tag{14}$$

其中：$c(s_i)$ 表示词义 $s_i$ 出现的次数，$c(s_i,v_k)$ 表示 $s_i$ 和上下文 $v_k$ 共同出现的次数。

## 3　实验结果与分析

实验环境为：操作系统为 Windows 10，CPU 为 Intel Core i5-3337U, 内存 4 GB, 分类平台为 Weka。使用中科院 ICTCALS 官方提供的 ICTCLAS 汉语分词系统进行分词，使用哈工大信息检索实验室通用词表去除停用词，使用哈工大信息检索实验室扩展同义词林，使用 Weka 平台的 KNN 分类器进行分类。为考察优化效果，使用查准率、查全率和 F1 测试值对实验结果进行评估。

查准率考察分类系统的分类准确性，划分到类 $c_j$ 中的文本是否真的属于类 $c_j$，如式(15)所示。

$$P_p = \frac{S_r}{S_a} \times 100\% \tag{15}$$

查全率考察分类系统是否将属于类 $c_j$ 的所有文本都划分到类 $c_j$，公式为

$$P_c = \frac{S_r}{S_o} \times 100\% \tag{16}$$

可用 F1 值来综合考察这两个指标，F1 值公式为

$$F1 = \frac{P_p \times P_c \times 2}{P_p + P_c} \times 100\% \tag{17}$$

比较经典 CHI 方法和优化后的 CHI 方法的三项指标数值，以及分类准确提升率。实验数据语料库为搜狗实验室提供的人工分类的中文文本语料库，涵盖教育、体育、军事、文化、经济、计算机、健康、工作、旅游等 9 个类别，从中选取每类 400 篇，共计 3600 篇文章，使用 4 折交叉验证方法，将语料库每类分成均等 4 份，每次实验取其中三份作为训练集，剩下的作为测试集，重复实验四次去平均值作为实验结果。实验时，首先使用 ICTCLAS 对文本进行分词，分词后根据 TF-IDF 公式计算文本的初试特征向量，保存该特征向量，保证之后的对比实验使用同一实验数据。

该实验分两部分，首先利用原始的 VSM 方法对文本进行建模后，利用特征项 TF-IDF 权值对其排序，从中选取出代表文本的特征向量；然后使用传统的 CHI 方法、IG 方法和优化有的 CHI 方法进行分类，测得三种方法分类的查全率、查准率和 F1 值。三种方法的查准率、查全率和 F1 值比较，如图 2~4 所示。

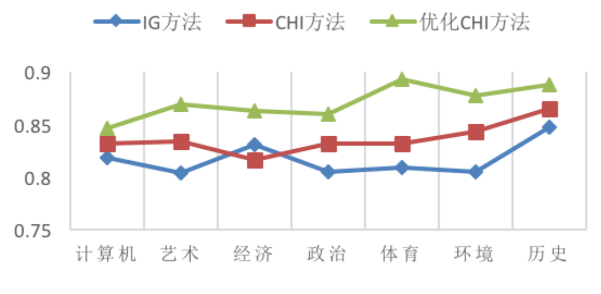



图 4　三种方法的查准率比较

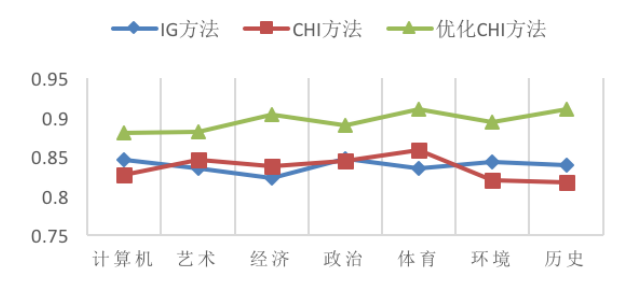



图 5　三种方法的查全率比较

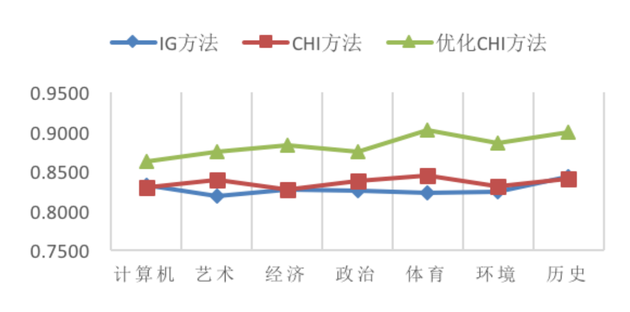



图 6　三种方法的 F1 值比较

从上述图表中可看出，同一算法对不同类别的文本的指标都略有不同，比如体育类的文本，其查准率相对来说都比较高，因为体育类的有许多专有名词。而经济、政治等类，有很多重合的部分，因此查准率相对较低。在查全率方面，三种算法在查全率方面都比较稳定，在不同的类别，IG 和传统 CHI 表现各有优劣，性能相差不大，优化后 CHI 方法则相比两种方法有明显改善。本文提出的 CHI 优化算法在文本分类中，相较传统 IG 方法和 CHI 方法在查全率方面提升效果明显，在查准率方面则不太稳定。具体实验结果如表 2 所示。对比优化后的 CHI 方法相比两种传统方法的 F1 值提升率如表 3、4 所示。

表 2　CHI 方法和优化 CHI 方法文本分类数据统计

| 类别 | | 计算机 | 艺术 | 经济 | 政治 | 体育 | 环境 | 历史 | 均值 |
|---|---|---|---|---|---|---|---|---|---|
| IG 方法 | 查准率 | 0.8182 | 0.8033 | 0.8309 | 0.8048 | 0.8093 | 0.8048 | 0.8477 | 0.8170 |
| | 查全率 | 0.8464 | 0.8352 | 0.8231 | 0.8478 | 0.8355 | 0.8436 | 0.8397 | 0.8388 |
| | F1 值 | 0.8321 | 0.8189 | 0.8270 | 0.8257 | 0.8222 | 0.8237 | 0.8437 | 0.8276 |
| CHI 方法 | 查准率 | 0.8323 | 0.8335 | 0.8162 | 0.8322 | 0.8313 | 0.8432 | 0.8655 | 0.8363 |
| | 查全率 | 0.8266 | 0.8458 | 0.8379 | 0.8442 | 0.8591 | 0.8199 | 0.8173 | 0.8358 |
| | F1 值 | 0.8294 | 0.8396 | 0.8269 | 0.8382 | 0.8450 | 0.8314 | 0.8407 | 0.8359 |
| 优化 CHI 方法 | 查准率 | 0.8466 | 0.8697 | 0.8629 | 0.8599 | 0.8927 | 0.8774 | 0.8883 | 0.8711 |
| | 查全率 | 0.8801 | 0.8812 | 0.9033 | 0.8903 | 0.9111 | 0.8935 | 0.9101 | 0.8957 |
| | F1 值 | 0.8630 | 0.8754 | 0.8826 | 0.8748 | 0.9018 | 0.8854 | 0.8991 | 0.8832 |

表 3　优化 CHI 方法相比 IG 方法的 F1 值提升率

| 类别 | 计算机 | 艺术 | 经济 | 政治 | 体育 | 环境 | 历史 | 均值 |
|---|---|---|---|---|---|---|---|---|
| IG F1 值 | 0.8321 | 0.8189 | 0.8270 | 0.8257 | 0.8222 | 0.8237 | 0.8437 | 0.8276 |
| 优化 CHI F1 值 | 0.8630 | 0.8754 | 0.8826 | 0.8748 | 0.9018 | 0.8854 | 0.8991 | 0.8832 |
| F1 值提升率 | 3.72% | 6.90% | 6.73% | 5.95% | 9.68% | 7.48% | 6.56% | 6.71% |

表 4　优化 CHI 方法相比 CHI 方法的 F1 值提升率

| 类别 | 计算机 | 艺术 | 经济 | 政治 | 体育 | 环境 | 历史 | 均值 |
|---|---|---|---|---|---|---|---|---|
| CHI F1 值 | 0.8294 | 0.8396 | 0.8269 | 0.8382 | 0.8450 | 0.8314 | 0.8407 | 0.8359 |
| 优化 CHI F1 值 | 0.8630 | 0.8754 | 0.8826 | 0.8748 | 0.9018 | 0.8854 | 0.8991 | 0.8832 |
| F1 值提升率 | 4.05% | 4.26% | 6.74% | 4.38% | 6.73% | 6.49% | 6.94% | 5.66% |

在同义词聚类时，一词多义可能会对实验结果带来影响。为此，增加一组实验对比未加多义词义项判别的同义词聚类算法和新增多义词判别的同义词聚类算法。由于新闻文章相较其他类型的文章，更贴合人类的自然表达，用词相对更为多变，同义词和多义词出现情况更为频繁，因此使用搜狗实验室提供的全网新闻数据语料库进行测试，该语料库包含财经、健康、文化等 10 个频道的新闻数据，总大小为 1.02GB。同样使用查准率、查全率和 F1 测试值对实验结果进行评估。实验数据如表 5、6 所示。

表 5　同义词聚类方法分类数据

| 属性 | 汽车 | 财经 | IT | 健康 | 体育 | 旅游 | 教育 | 招聘 | 文化 | 军事 | 均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 查准率 | 0.8425 | 0.8257 | 0.8633 | 0.8526 | 0.8439 | 0.835 | 0.8024 | 0.8222 | 0.8017 | 0.8556 | 0.83449 |
| 查全率 | 0.8377 | 0.8004 | 0.8331 | 0.8024 | 0.8426 | 0.8131 | 0.8335 | 0.8355 | 0.8152 | 0.823 | 0.82365 |
| F1 值 | 0.8401 | 0.8129 | 0.8479 | 0.8267 | 0.8432 | 0.8239 | 0.8177 | 0.8288 | 0.8084 | 0.8390 | 0.82886 |

表 6　多义词优化方法分类数据

| 属性 | 汽车 | 财经 | IT | 健康 | 体育 | 旅游 | 教育 | 招聘 | 文化 | 军事 | 均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 查准率 | 0.8601 | 0.836 | 0.8707 | 0.8755 | 0.8631 | 0.8777 | 0.8459 | 0.8507 | 0.8559 | 0.8747 | 0.86103 |
| 查全率 | 0.8410 | 0.8136 | 0.8451 | 0.8312 | 0.8437 | 0.825 | 0.8437 | 0.8519 | 0.8204 | 0.8301 | 0.83457 |
| F1 值 | 0.8504 | 0.8246 | 0.8577 | 0.8528 | 0.8533 | 0.8505 | 0.8448 | 0.8513 | 0.8378 | 0.8518 | 0.84751 |

采用多义词义项判别后的同义词聚类方法比简单同义词聚类方法在查准率上有明显提升。两者性能比较如图 7~9 所示。

从上述图表中可看出，使用增加多义词词义判别后，同义词聚类的方法性能有明显提升。主要是在查准率方面比简单同义词聚类方面有了明显改善，特别是在教育、文化等类别，这些类别一词多义的情况可能相对较多，简单同义词聚类方法没有考虑一词多义情况，可能导致分类错误。使用多义词义项判别优化的同义词聚类方法后，减低了一词多义情况对分类结果的影响，提高了查准率。同义词聚类方法相比传统方法查全率有明显改善，而针对一词多义提出的多义词义项判别则在同义词聚类的基础上提高了查准率。

## 4　结束语

文本分类中，最重要的是准确提取文本特征项和降维。特征项的准确提取最重要的是准确评估每个特征项的权值，降维最重要的是选择性剔除噪声数据并提高文本文本特征提取信噪比。文本针对传统文本分类方法中的卡方统计方法进行了改进，通过同义词聚类合并词义相近的特征项，降低了维数；针对卡方统计对低频次敏感的特点，提出了基于词频的特征项权值改进。在降低维数的同时，提高特征项选择的准确性。对简单同义词聚类可能出现的一词多义情况，提出基于多义词义项判别的同义词聚类优化方法，实验结果表明，新的方法在查准率、查全率、F1 值等方面相比传统的 CHI 和 IG 有很大提升。

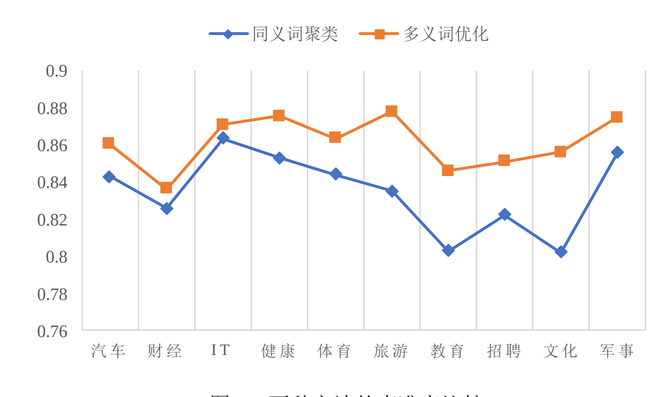



图 7　两种方法的查准率比较

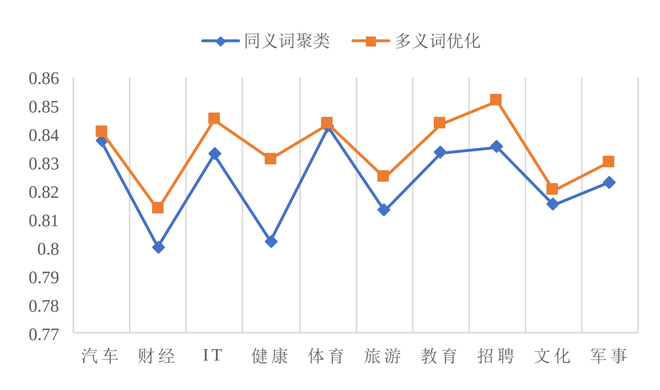



图 8 两种方法的查全率比较

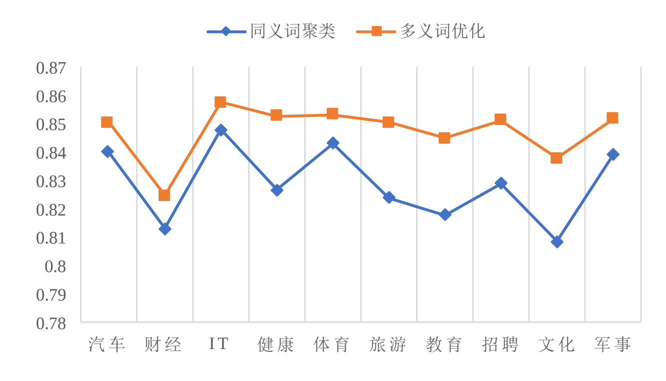



图 9 两种方法的 F1 值比较

## 参考文献：


[1] 毛晓刚. 基于向量空间模型的本地搜索引擎的设计与实现 [D]. 长春: 吉林大学, 2016.

[2] 刘海峰, 苏展, 刘守生. 一种基于词频信息的改进 CHI 文本特征选择 [J]. 计算机工程与应用, 2013, 49 (22): 110-114.

[3] Akce A, Norton J J S, Bretl T. An SSVEP-based brain–computer interface for text spelling with adaptive queries that maximize information gain rates [J]. IEEE Trans on Neural Systems & Rehabilitation Engineering, 2015, 23 (5): 857-866.

[4] Yin C, Xiang J, Zhang H, *et al.* A new SVM method for short text classification based on semi-supervised learning [C]// Proc of International Conference on Advanced Information Technology and Sensor Application. 2016: 100-103.

[5] Rao Y, Xie H, Li J, *et al.* Social emotion classification of short text via topic-level maximum entropy model [J]. Information & Management, 2016, 53 (8): 978-986.

[6] Han Y, Li Meicong, Guo X C, *et al.* The Text classification attribute reduction algorithm based on the rough set theory [J]. Journal of Northeast Dianli University, 2016.

[7] Rashmi S, Hanumanthappa M, Reddy M V. Hidden Markov model for speech recognition system: a pilot study and a naive approach for speech-to-text model [M]// Speech and Language Processing for Human-Machine Communications. Advances in Intelligent Systems and Computing. 2017: 77-90.

[8] Kenter T, Rijke M D. Short Text Similarity with Word Embeddings [C]// Proc of ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1411-1420.

[9] Song Y, Dan R. Unsupervised Sparse Vector Densification for Short Text Similarity [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1275-1280.

[10] Cheng X. A study on lexical sense relations from the perspective of vocabulary breadth and word frequency [J]. Theory & Practice in Language Studies, 2016, 6 (5): 988.

[11] Khan M I. The investigation and importance of sense-relations and semantics in the English language [J]. Language in India, 2016, 16 (3): 106.

[12] Chiang H H, Lee T S. Family relations, sense of coherence, happiness and perceived health in retired taiwanese: analysis of a conceptual model [J]. Geriatrics & Gerontology International, 2018, 18 (1): 154-160.

[13] 周庆平, 谭长庚, 王宏君, 等. 基于聚类改进的 KNN 文本分类算法 [J]. 计算机应用研究, 2016, 33 (11): 3374-3377.

[14] 裴英博, 刘晓霞. 文本分类中改进型 CHI 特征选择方法的研究 [J]. 计算机工程与应用, 2011, 47 (4): 128-130.

[15] 熊忠阳, 张鹏招, 张玉芳. 基于χ2 统计的文本分类特征选择方法的研究 [J]. 计算机应用, 2008, 28 (2): 513-514.

[16] 王光, 邱云飞, 史庆伟. 集合 CHI 与 IG 的特征选择方法 [J]. 计算机应用研究, 2012, 29 (7): 2454-2456.

[17] 邱云飞, 王威, 刘大有, 等. 基于方差的 CHI 特征选择方法 [J]. 计算机应用研究, 2012, 29 (4): 1304-1306.